Journal of Multivariate Analysis **67**, 277–296 (1998)
Article No. MV981765

Prediction and Classification of Non-stationary Categorical Time Series*

Konstantinos Fokianos†

The Ohio State University

and

Benjamin Kedem

University of Maryland

Received February 14, 1996; revised June 16, 1998

stationary categorical time series is presented, taking into account stochastic time dependent covariates. The model links the probabilities of each category to a covariate process through a vector of time invariant parameters. Under mild regularity conditions, we establish good asymptotic properties of the estimator by appealing to martingale theory. Certain diagnostic tools are presented for checking the adequacy of the fit. © 1998 Academic Press

AMS 1991 subject classifications: primary, 62M10; secondary, 62F12.

Key words and phrases: non-stationarity; classification; prediction; asymptotic theory; partial likelihood; goodness of fit.

1. INTRODUCTION

Categorical time series arise in numerous applications, many of which are reported in the recent books by Diggle, Liang, and Zeger [9], Fahrmeir and Tutz [11], and Kedem [16, Ch. 9]. Examples of categorical time series include signals quantized at several levels, clipped binary time series, and any multi-response longitudinal data observed on an ordinal or nominal scale. And just as with "ordinary" time series the problem of forecasting or prediction in categorical series is of importance, except that usually it concerns the estimation of a future transition probability given past data and auxiliary information. In this regard, the prediction problem is essentially synonymous with the problem of classification of a future

* Work supported by grants NSF EEC-940-2384, ONR N00014-92-C-0019, NASA NAG 5-2783.

† Present address: Dept. of Mathematics and Statistics, University of Cyprus, P.O. Box 537, 1678 Nicosia, Cyprus.

0047-259X/98 $25.00

value in one of several categories given the past. In some cases, the complete dependence structure is known thus making statistical inference relatively easy. For example, when the series can be regarded as a homogeneous Markov chain, the inference problem can be attacked using the methods of Billingsley [4], but when the complete dependence structure is unknown, the problem becomes quite challenging.

Recent advances in categorical time series owe greatly to the introduction of generalized linear models and *link* functions as described in McCullagh and Nelder [22]. Accordingly, the one step transition probability is conveniently parametrized via the link, and this goes along well with conditional inference, allowing for some form of non-stationarity. Conditional inference where a Markov assumption is made can be found in Korn and Whittemore [18], Keenan [17], Stern and Coe [30], Muenz and Rubinstein [23], Bonney [5], Fahrmeir and Kaufmann [10], Kaufmann [14], Liang and Zeger [21], Li [20], Brillinger [7] to name only a few.

In particular, we mention that asymptotic theory for regression models for non-stationary categorical time series has been previously studied by Fahrmeir and Kaufmann [10] and Kaufmann [14]. However, our work extends their results by considering *random time dependent covariates* and by dropping any markovian assumption. In addition, motivated by the work of Slud and Kedem [29]—who only dealt with logistic regression for binary time series—we use an ergodicity assumption which might be useful for other problems and introduce a goodness of fit test statistic. We perform conditional inference using *partial likelihood*, a concept introduced by Cox [8], and extended and ramified by Wong [32], Slud [28]. Partial likelihood simplifies conditional inference—for example, it obviates the Markov assumption—and is particularly useful for time series where the dependence is unknown, let alone the knowledge of joint distributions. Furthermore, as noted by Murphy and Li [24], partial likelihood inference allows missing values. Indeed, all that is needed is a conditional model relative to a nested sequence of histories.

Following Wong [32] and Slud [28], we first give the definition of partial likelihood, and then setup the model. We next discuss the large sample theory. This is followed by some diagnostic tools.

2. THE MATHEMATICAL SETUP

2.1. *Partial Likelihood*

Assume that an individual observes a stochastic process, say (x_t, y_t), $t = 1, ..., N$. In principle, we can write down the joint distribution of all the

observations up to time N, by employing the law of total probability; that is (Wong [32])

$$f(x_1, y_1, x_2, y_2, ..., x_N, y_N) = \left[\prod_{t=1}^{N} f(y_t \mid d_t) \right] \left[\prod_{t=1}^{N} f(x_t \mid c_t) \right], \qquad (1)$$

where $d_t = (y_1, x_1, ..., y_{t-1}, x_{t-1})$ and $c_t = (y_1, x_1, ..., y_{t-1}, x_{t-1}, y_t)$.

Cox [8] defined the second product on the right hand side of (1) as the *Partial Likelihood*. It is helpful to note that the σ-filed generated by c_{t-1} is contained in the one generated by c_t. This is a key feature which motivates our definition (see Slud [28], Slud and Kedem [29]).

DEFINITION 2.1. Let $\mathcal{F}_t$, $t = 0, 1, ...$ be an increasing sequence of σ-fields, $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 ...$, and let $X_1, X_2, ...$ be a sequence of random variables on some common probability space such that X_t is $\mathcal{F}_t$ measurable. Denote the density of X_t given $\mathcal{F}_{t-1}$ by $f_t(x_t; \beta)$, where $\beta \in R^p$ is a parameter. The partial likelihood (PL) function relative to β, $\mathcal{F}_t$, and the data $X_1, X_2, ..., X_N$, is given by the product

$$PL(\beta; X_1, ..., X_N) = \prod_{t=1}^{N} f_t(x_t; \beta). \qquad (2)$$

This definition generalizes both likelihood and conditional likelihood. Unlike (full) likelihood, partial likelihood does not require complete knowledge of the joint distribution of the covariates. Unlike conditional likelihood, complete covariate information need not be known throughout the period of observation. Partial likelihood takes into account only what is known to the observer up to the time of actual observation.

The vector β that maximizes (2) is called the maximum partial likelihood estimator (MPLE). Its asymptotic distribution has been studied by several authors (see Wong [32]; Slud and Kedem [29]). In the context of survival analysis and counting processes see Andersen and Gill [2], Arjas and Haara [3], for example. The key point is that the gradient of the logarithm of (2) is a martingale with respect to the nested sequence of histories $\mathcal{F}_t$.

2.2. *A General Model*

We introduce now some notation and terminology which will be found useful in the sequel. Suppose that we actually observe a non-stationary categorical time series, say $\{ y_t, t = 0, 1, ..., N \}$. Let m denote the number of possible categories and assume that the $t'th$ observation is given by a vector $y_t = (y_{t1}, ..., y_{tq})'$ of length $q = m - 1$, where

$$y_{tj} = \begin{cases} 1 & \text{if the } j\text{th category is observed at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

Let $p_t = (p_{t1}, ...p_{tq})'$ denote the corresponding vector of conditional probabilities given $\mathscr{F}_{t-1}$. In other words $p_{tj} = P(y_{tj} = 1 \parallel \mathscr{F}_{t-1})$ for $j = 1, ..., q$. The σ-algebra $\mathscr{F}_{t-1}$ represents the whole available information to the observer up to and including time t. For the $m'th$ category, put

$$y_{tm} = 1 - \sum_{j=1}^{q} y_{tj} \tag{3}$$

and

$$p_{tm} = 1 - \sum_{j=1}^{q} p_{tj}. \tag{4}$$

Assume that $\mathbf{Z}_{t-1}$ is a $p \times q$ matrix that represents a covariate process. In other words each response y_t corresponds to a time-dependent random covariate matrix $\mathbf{Z}_{t-1}$. The covariate matrix usually consists of any lagged values of the response process and (or) any exogenous variables that evolve in time simultaneously with the response variable. Moreover, lagged values of the exogenous variables are allowed as well as any interactions between the response and the covariates.

The aim of this paper is to develop an asymptotic theory for a flexible and parsimonious class of models that *link* the probability of the jth category with the covariate process in a certain way. This leads to an attractive parametrization, which extends ideas from the generalized linear models (GLIM) and the autoregressive moving average models (ARMA) (box and Jenkins [6]).

Define (see Fahrmeir and Kaufmann [10], Kaufmann [14])

$$p_t = h(\mathbf{Z}'_{t-1}\beta). \tag{5}$$

Here β denotes a p dimensional vector of time invariant unknown parameters which belongs to an open set $B \subseteq R^p$. The function h is called the *link function*. We assume that the link function maps a subset $H \subseteq R^q$ bijectively onto $\{(w_1, ..., w_q)'\colon\ w_j > 0,\ j = 1, ..., q,\ \sum_{j=1}^{q} w_j < 1\}$. Note that the multinomial logits and the cumulative odds models fall in this category (Agresti [1]).

In our context we have the multinomial probability

$$f(y_t; \beta \mid \mathscr{F}_{t-1}) = \prod_{j=1}^{m} p_{tj}(\beta)^{y_{tj}}. \tag{6}$$

Consequently, the corresponding Partial Likelihood is

$$PL(\beta) = \prod_{t=1}^{N} f(y_t; \beta \mid \mathcal{F}_{t-1})$$
$$= \prod_{t=1}^{N} \prod_{j=1}^{m} p_{tj}(\beta)^{y_{tj}}. \tag{7}$$

It follows that the partial log-likelihood is given by

$$pl_N(\beta) = \sum_{t=1}^{N} \sum_{j=1}^{m} y_{tj} \log p_{tj}(\beta). \tag{8}$$

The partial score is given by the vector

$$S_N(\beta) = \left(\frac{\partial pl_N(\beta)}{\partial \beta_1}, ..., \frac{\partial pl_N(\beta)}{\partial \beta_p} \right)'. \tag{9}$$

It follows that

$$S_N(\beta) = \sum_{t=1}^{N} \mathbf{Z}_{t-1} \mathbf{D}_{t-1}(\beta)(y_t - p_t(\beta)), \tag{10}$$

where $\mathbf{D}_{t-1}(\beta) = [\partial d(\mathbf{Z}'_{t-1}\beta)/\partial \gamma_{t-1}]$ with $\gamma_{t-1} = \mathbf{Z}'_{t-1}\beta$. The function d is defined as the composition of the functions h and l with l standing for the logits function. This function is defined by

$$l(p_t) = (\log(p_{t1}/p_{tm}), ..., \log(p_{tq}/p_{tm})).$$

The conditional information matrix is given by

$$\mathbf{G}_N(\beta) = \sum_{t=1}^{N} Cov[\mathbf{Z}_{t-1} \mathbf{D}_{t-1}(\beta)(y_t - p_t(\beta)) \mid \mathcal{F}_{t-1}]$$
$$= \sum_{t=1}^{N} \mathbf{Z}_{t-1} \mathbf{D}_{t-1}(\beta) \mathbf{\Sigma}_t \mathbf{D}'_{t-1}(\beta) \mathbf{Z}'_{t-1} \tag{11}$$

with $\mathbf{\Sigma}_t(\beta)$ is the conditional covariance matrix of y_t with generic element

$$\sigma_t^{(ij)}(\beta) = \begin{cases} -p_{ti}(\beta) p_{tj}(\beta) & \text{if} \quad i \neq j \\ p_{ti}(\beta)(1 - p_{ti}(\beta)) & \text{if} \quad i = j \end{cases}$$

for $i, j = 1..., q$. The unconditional information matrix is

$$\mathbf{F}_N(\beta) = E[\mathbf{G}_N(\beta)]. \tag{12}$$

Finally, the second derivative of the partial log likelihood multiplied by -1, is

$$\mathbf{H}_N(\beta) = -\frac{\partial^2 pl_N(\beta)}{\partial\beta\,\partial\beta'} = \mathbf{G}_N(\beta) - \mathbf{R}_N(\beta), \tag{13}$$

where

$$\mathbf{R}_N(\beta) = \sum_{t=1}^{N}\sum_{r=1}^{q} \mathbf{Z}_{t-1}\mathbf{W}_{(t-1)r}(\beta)\,\mathbf{Z}'_{t-1}(y_{tr} - p_{tr}(\beta))$$

with $\mathbf{W}_{(t-1)r}(\beta) = [\partial^2 d_r(\mathbf{Z}'_{t-1}\beta)/\partial\gamma_{t-1}\,\partial\gamma'_{t-1}]$. Notice that the above expectation and variance are taken with respect to the true parameter. The maximum partial likelihood estimator (MPLE) is the consistent solution of the $S_N(\beta) = 0$. We point out that existence and uniqueness of the estimator is not guaranteed for finite samples. However, we obtain concavity of the log-likelihood for many important applications. In the setting of independent observation these questions have been studied by several authors. Among them are Haberman [12], Wedderburn [31], Pratt [25], Silvapulle [27], Kaufmann [15].

3. LARGE SAMPLE THEORY

We prove now existence, consistency and asymptotic normality of the MPLE under regularity conditions. We will consistently suppress any notation which depend on the true parameter.

Assumption (*A*). A.1. The parameter β belongs to an open set $B \subseteq R^p$.

A.2. The covariate matrix $\mathbf{Z}_{t-1}$ almost surely lies in a nonrandom compact subset Γ of $R^{p\times q}$ such that $P[\sum_{t=1}^{N}\mathbf{Z}_{t-1}\mathbf{Z}'_{t-1} > 0] = 1$. Furthermore we assume that $\mathbf{Z}'_{t-1}\beta$ lies almost surely in the domain H of h of for all $\mathbf{Z}_{t-1} \in \Gamma$ and $\beta \in B$.

A.3. The probability measure P which governs $\{y_t, \mathbf{Z}_{t-1}\}$, $t = 1, ..., N$, obeys (5) with $\beta = \beta_0$.

A.4. The link function h is twice continuously differentiable, $\det[\partial h(\gamma)/\partial\gamma] \neq 0$.

A.5. There is a probability measure μ on $R^{p\times q}$ such that $\int_{R^{p\times q}} \mathbf{Z}\mathbf{Z}'\mu(d\mathbf{Z})$ is positive definite, and such that under (5) with $\beta = \beta_0$, for Borel sets $A \subset R^{p\times q}$ we have

$$\frac{1}{N}\sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1}\in A]} \xrightarrow{p} \mu(A), \qquad \text{as} \quad N \to \infty.$$

Note that integration with respect to a matrix means that we integrate with respect to each element of the matrix. Assumptions A.1 and A.4 guarantee that the second derivative of the partial log-likelihood is a continuous function of β. The condition $\det[\partial h(\gamma)/\partial\gamma] \neq 0$ implies in particular that $\mathbf{D}_{t-1}$ is not singular, so from A.2 the conditional information matrix is positive definite with probability 1. To see this note that for any vector $\lambda \in R^p$, $\lambda \neq 0$,

$$\begin{aligned}\lambda' \mathbf{G}_N \lambda &= \lambda' \left(\sum_{t=1}^{N} \mathbf{Z}_{t-1} \mathbf{D}_{t-1} \boldsymbol{\Sigma}_t \mathbf{D}'_{t-1} \mathbf{Z}'_{t-1} \right) \lambda \\ &\geqslant \min_t \lambda_{\min}(\mathbf{D}_{t-1} \boldsymbol{\Sigma}_t \mathbf{D}'_{t-1}) \left(\lambda' \sum_{t=1}^{N} \mathbf{Z}_{t-1} \mathbf{Z}'_{t-1} \lambda \right), \\ &> 0 \qquad \text{a.s.}\end{aligned}$$

where $\lambda_{\min}$ denotes the minimum eigenvalue. Since the variance-covariance matrix is positive definite and the matrix of derivatives, $\mathbf{D}_{t-1}$, is not singular we have that the minimum eigenvalue is positive almost everywhere. It follows that the unconditional information matrix is positive definite as well. The last part of assumption A.2 assures that we have a well defined model. The compactness assumption will be useful in deriving bounds for the asymptotics. Assumption A.5 simply states that if g is any continuous and bounded function on Γ taking values on $R^{p\times q}$ then we have that

$$\frac{\sum_{t=1}^{N} g(\mathbf{Z}_{t-1})}{N} \xrightarrow{p} \int_{R^{p\times q}} g(\mathbf{Z})\, \mu(d\mathbf{Z}).$$

Thus the conditional information matrix $\mathbf{G}_N(\beta)$ has a non-random limit

$$\frac{\mathbf{G}_N(\beta)}{N} \xrightarrow{p} \int_{R^{p\times q}} \mathbf{Z}\mathbf{D}(\beta)\, \boldsymbol{\Sigma}(\beta)\, \mathbf{D}'(\beta)\, \mathbf{Z}' \mu(d\mathbf{Z}) = \mathbf{G}(\beta) \tag{14}$$

where $\mathbf{D}(\beta) = [\partial h(\mathbf{Z}'\beta)/\partial(\mathbf{Z}'\beta)]$ and $\boldsymbol{\Sigma}$ has generic element

$$\sigma^{(ij)}(\beta) = \begin{cases} -h_i(\mathbf{Z}'\beta)\, h_j(\mathbf{Z}'\beta) & \text{if} \quad i \neq j \\ h_i(\mathbf{Z}'\beta)(1 - h_i(\mathbf{Z}'\beta)) & \text{if} \quad i = j \end{cases}$$

for $i, j = 1..., q$. From (A.5), $\mathbf{G}(\beta)$ is a positive definite matrix at the true value and therefore its inverse exists. It is important to emphasize that our approach is quite general and does not call for any Markov assumption (compare with Fahrmeir and Kaufmann [10]; Kaufmann [14]).

At this point, we want to mention that we will use the right Cholesky square root of a positive definite matrix in the sequel. More precisely, if $\mathbf{B}$

is a positive definite matrix then the right Cholesky square root, denoted by $\mathbf{B}^{1/2}$, is defined as the unique upper triangular matrix with positive elements such that $\mathbf{B}=(\mathbf{B}^{1/2})'\,(\mathbf{B}^{1/2})$. We denote by $\mathbf{B}^{t/2}=(\mathbf{B}^{1/2})'$. Our proof of consistency and asymptotic normality is based on the classical approach, namely we first exhibit a solution of the score equations and then prove that is consistent and asymptotically normally distributed.

To this end, we will need some helpful lemmas. The following lemma shows that the partial score process is a zero mean square integrable martingale which satisfies the conditions for an application of a martingale central limit theorem.

LEMMA 3.1. *Consider the model* (5) *and assume that assumption* (*A*) *holds. Then the partial score process* $\{S_t, \mathscr{F}_t\}$ *is a zero mean square integrable martingale such that*:

$$\mathbf{F}_N^{-1/2} S_N \xrightarrow{D} \mathscr{N}$$

as $N \to \infty$, *with* $\mathscr{N}$ *denoting a standard normal random vector.*

Proof. The fact that the partial score process is zero mean square integrable martingale follows from (10) and assumption A.2. To show it actually converges in distribution, we consider $\phi_N=\lambda' S_N$, with $\lambda \in R^p$, having in mind the Cramér–Wold device. Then ϕ_N is a univariate zero-mean martingale. Its conditional and unconditional covariance matrices are $\lambda' \mathbf{G}_N \lambda$ and $\lambda' \mathbf{F}_N \lambda$ respectively. Thus

$$\frac{\lambda' \mathbf{G}_N \lambda}{\lambda' \mathbf{F}_N \lambda} = \frac{\lambda' \mathbf{G}_N \lambda / N}{\lambda' \mathbf{F}_N \lambda / N} \xrightarrow{p} \frac{\lambda' \mathbf{G} \lambda}{\lambda' \mathbf{G} \lambda} = 1$$

upon invoking A.2 and A.5. Furthermore, by letting $I_{Nt}(\varepsilon)$ to be the indicator of the set $\{|\lambda' a_t|^2 \geqslant (\lambda' \mathbf{F}_N \lambda)^{1/2} \varepsilon\}$ with $a_t = S_t - S_{t-1}$, we get

$$\begin{aligned} \frac{1}{\lambda' \mathbf{F}_N \lambda} \sum_{t=1}^{N} E[\,|\lambda' a_t|^2 I_{Nt}(\varepsilon) \parallel \mathscr{F}_{t-1}] &\leqslant \frac{1}{(\lambda' \mathbf{F}_N \lambda)^{3/2} \varepsilon} \sum_{t=1}^{N} E[\,|\lambda' a_t|^3 \parallel \mathscr{F}_{t-1}] \\ &\leqslant \frac{N M_1}{(\lambda' \mathbf{F}_N \lambda)^{3/2} \varepsilon} \end{aligned}$$

where M_1 is a bound. Such a bound exists from A.2. Therefore Lindeberg's condition holds since the right hand side of the above tends to zero. The conclusion of the lemma follows from the central limit theorem for martingales (Hall and Heyde [13, Corollary 3.1]). ∎

The next lemma, a consequence of the Linderberg's condition, parallels the well-known result from linear models (Lai and Wei [19]).

LEMMA 3.2. *Under* **(A)** *we have that*

$$\lambda_{\min}(\mathbf{F}_N) \to \infty$$

as $N \to \infty$, *where* $\lambda_{\min}$ *is the minimum eigenvalue of the unconditional information matrix.*

Proof. Recall that if **A** and **B** are positive definite matrices,

$$|\lambda_{\min}(\mathbf{A}) - \lambda_{\min}(\mathbf{B})| \leqslant c \, \|\mathbf{A} - \mathbf{B}\| \tag{15}$$

where the positive constant depends only on the norm of the matrix. Then, by the proof of Lemma 3.1, we get

$$\left| \lambda_{\min}\left(\frac{\mathbf{F}_N}{N}\right) - \lambda_{\min}(\mathbf{G}) \right| \to \mathbf{0}.$$

It follows that $\lambda_{\min}(\mathbf{F}_N) = \bigcirc(N)$ and the claim is true. ▌

We prove now a continuity condition. Namely, we would like to have the matrix of second derivatives as close as possible to the information matrix. This is a technical lemma and the proof is along the lines of Kaufmann [14].

LEMMA 3.3. *Under* **(A)** *the following continuity condition holds*

$$\sup_{\tilde{\beta} \in O_N(\delta)} \|\mathbf{F}_N^{-1/2}(\mathbf{H}_N(\tilde{\beta}) - \mathbf{G}_N)\, \mathbf{F}_N^{-t/2}\| \xrightarrow{p} 0$$

with $O_N(\delta) = \{\tilde{\beta}\colon \|\mathbf{F}_N^{t/2}(\tilde{\beta} - \beta)\| \leqslant \delta\}$, *for any* $\delta > 0$ *and any matrix norm.*

Proof. Let $\lambda \in R^p$, with $\lambda \neq 0$ and assume without loss of generality that $\|\lambda\| = 1$. We will show the equivalent condition, for any $\delta > 0$

$$\sup_{\tilde{\beta} \in O_N(\delta)} \lambda' \mathbf{F}_N^{-1/2}(\mathbf{H}_N(\tilde{\beta}) - \mathbf{G}_N)\, \mathbf{F}_N^{-t/2}\lambda \xrightarrow{p} 0 \tag{16}$$

using once more Cramér–Wold device. By decomposing $\mathbf{H}_N(\tilde{\beta}) = \mathbf{G}_N(\tilde{\beta}) - \mathbf{R}_N(\tilde{\beta})$ we need really to show

$$g_N = \sup_{\tilde{\beta} \in O_N(\delta)} \lambda' \mathbf{F}_N^{-1/2}(\mathbf{G}_N(\tilde{\beta}) - \mathbf{G}_N)\, \mathbf{F}_N^{-t/2}\lambda \xrightarrow{p} 0 \tag{17}$$

and

$$\sup_{\tilde{\beta} \in O_N(\delta)} \lambda' \mathbf{F}_N^{-1/2} \mathbf{R}_N(\tilde{\beta})\, \mathbf{F}_N^{-t/2}\lambda \xrightarrow{p} 0 \tag{18}$$

hold simultaneously. Define the vectors $w'_{(t-1)N} = \lambda' \mathbf{F}_N^{-1/2} \mathbf{Z}_{t-1}$, for $1 \leqslant t \leqslant N$, and $w_N = \sum_{t=1}^N w'_{(t-1)N} w_{(t-1)N}$. Then we have that

$$g_N = \sup_{\tilde{\beta} \in O_N(\delta)} \sum_{t=1}^N w'_{(t-1)N} (\mathbf{L}_{t-1}(\tilde{\beta}) - \mathbf{L}_{t-1})\, w_{(t-1)N},$$

where $\mathbf{L}_{t-1}(\beta) = \mathbf{D}_{t-1} \boldsymbol{\Sigma}_t \mathbf{D}'_{t-1}$ for $t = 1, ..., N$. It follows that

$$g_N \leqslant w_N \sup_{\tilde{\beta} \in O_N(\delta),\, t} \|\mathbf{L}_{t-1}(\tilde{\beta}) - \mathbf{L}_{t-1}\|.$$

Using A.2, $\sup_t \|\mathbf{L}_{t-1}(\tilde{\beta}) - \mathbf{L}_{t-1}\|$ can be estimated from above by a continuous function of $\tilde{\beta}$ with a zero at $\tilde{\beta} = \beta$. Notice that $\{O_N(\delta)\}$ shrinks to β. Hence

$$\sup_{\tilde{\beta} \in O_N(\delta),\, t} \|\mathbf{L}_{t-1}(\tilde{\beta}) - \mathbf{L}_{t-1}\| \to 0.$$

By applying Markov's inequality we have that

$$\begin{aligned} P[|g_N| \geqslant \varepsilon] &\leqslant \frac{E[|g_N|]}{\varepsilon} \\ &\leqslant E[|w_N|] \sup_{\tilde{\beta} \in O_N(\delta),\, t} \|\mathbf{L}_{t-1}(\tilde{\beta} - \mathbf{L}_{t-1}\| \\ &\leqslant M \sup_{\tilde{\beta} \in O_N(\delta),\, t} \|\mathbf{L}_{t-1}(\tilde{\beta}) - \mathbf{L}_{t-1}\| \to 0, \end{aligned}$$

where M is a bound for w_N. Such a bound exists because of assumption (A.2) and the convergence of $\mathbf{F}_N$. By further decomposition we obtain

$$\sup_{\tilde{\beta} \in O_N(\delta)} \sum_{t=1}^N w'_{(t-1)N} (\mathbf{W}_{(t-1)j}(\tilde{\beta}) - \mathbf{W}_{(t-1)j})\, w_{(t-1)N} (y_{tj} - p_{tj}) \xrightarrow{p} 0 \qquad (19)$$

$$\sup_{\tilde{\beta} \in O_N(\delta)} \sum_{t=1}^N w'_{(t-1)N} \mathbf{W}_{(t-1)j}(\tilde{\beta})\, w_{(t-1)N} (p_{tj} - p_{tj}(\tilde{\beta})) \xrightarrow{p} 0 \qquad (20)$$

$$\sum_{t=1}^N w'_{(t-1)N} \mathbf{W}_{(t-1)j} w_{(t-1)N} (y_{tj} - p_{tj}) \xrightarrow{p} 0 \qquad (21)$$

for any j, $1 \leqslant j \leqslant q$, jointly are sufficient for (18). The proofs of (19) and (20), are the same as that of (17). To prove (21), consider the increments of (21), that is

$$u_{(t-1)N} = w'_{(t-1)N} \mathbf{W}_{(t-1)j} w_{(t-1)N} (y_{tj} - p_{tj}).$$

Then we see that

$$E[u_{(t-1)N} \| \mathscr{F}_{t-1}] = 0$$

and

$$\begin{aligned} Var[u_{(t-1)N} \| \mathscr{F}_{t-1}] &= w'_{(t-1)N} \mathbf{W}_{(t-1)j} w_{(t-1)N} \\ &\quad \times Var[y_{tj} - p_{tj} \| \mathscr{F}_{t-1}]\, w'_{(t-1)N} \mathbf{W}'_{(t-1)j} w_{(t-1)N} \\ &\leqslant K(w'_{(t-1)N} w_{(t-1)N})^2, \end{aligned}$$

where K is a bound on $\|\mathbf{W}_{(t-1)j}\|^2\, Var[y_{tj} - p_{tj} \| \mathscr{F}_{t-1}]$. Actually, the above two relations make clear that $\{u_{(t-1)N}, t = 1, ..., N\}$ are the orthogonal increments of a square integrable zero mean martingale. It follows that

$$E\left(\sum_{t=1}^{N} u_{(t-1)N} \right) = 0$$

and

$$\begin{aligned} Var\left[\sum_{t=1}^{N} u_{(t-1)N} \right] &\leqslant K \sum_{t=1}^{N} E[(w'_{(t-1)N} w_{(t-1)N})^2] \\ &\leqslant K \sup_t E[w'_{(t-1)N} w_{(t-1)N}]\, E[w_N]. \end{aligned}$$

However

$$\begin{aligned} \sup_t E(w'_{(t-1)N} w_{(t-1)N}) &= \sup_t \lambda' \mathbf{F}_N^{-1/2} E(\mathbf{Z}_{t-1} \mathbf{Z}'_{t-1})\, \mathbf{F}_N^{-t/2} \lambda \\ &\leqslant \lambda' \mathbf{F}_N^{-1} \lambda \sup_{\mathbf{Z}_{t-1} \in \Gamma} \|E(\mathbf{Z}_{t-1})\|^2 \\ &\leqslant \frac{\sup_{\mathbf{Z}_{t-1} \in \Gamma} \|E(\mathbf{Z}_{t-1})\|^2}{\lambda_{\min}(\mathbf{F}_N)} \to 0. \end{aligned}$$

Since $E[w_N]$ is bounded, from its convergence, relation (21) holds and therefore the continuity condition was established. ▌

We will need one more lemma before we prove our main results. Its proof is omitted.

LEMMA 3.4. *Suppose that* $\{X_N\}_{N=1}^{\infty}$ *and* $\{Y_N\}_{N=1}^{\infty}$ *are both sequences of positive random variables. In addition, assume that* $Y_N \xrightarrow{p} c$, *as* $N \to \infty$, *with* c *denoting a constant. Then*

$$|P(X_N \leqslant Y_N) - P(X_N \leqslant c)| \to 0$$

as $N \to \infty$.

We prove now the main result of this section.

THEOREM 3.1. *Under* **(A)**, *the probability that a locally unique maximum partial likelihood estimator exists converges to one. Moreover there exists a sequence of maximum partial likelihood estimators* $\hat{\beta}_N$ *which is consistent and asymptotically normal*:

$$\sqrt{N}(\hat{\beta}_N - \beta_0) \xrightarrow{D} \mathcal{N}(0, \mathbf{G}^{-1}(\beta_0)).$$

Proof. From Lemma 3.1 we get that

$$(\mathbf{F}_N^{-1/2} S_N, \mathbf{F}_N^{-1/2} \mathbf{G}_N \mathbf{F}_N^{-1/2}) \xrightarrow{D} (\mathcal{N}, \mathbf{I}),$$

where $\mathcal{N}$ is a standard normal vector. Choosing $\mathbf{G}_N^{1/2}$ such that $\mathbf{F}_N^{-1/2} \mathbf{G}_N^{1/2}$ is the Cholesky square root of $\mathbf{F}_N^{-1/2} \mathbf{G}_N \mathbf{F}_N^{-1/2}$, we have from the continuity of the square root that

$$(\mathbf{F}_N^{-1/2} S_N, \mathbf{F}_N^{-1/2} \mathbf{G}_N^{1/2}) \xrightarrow{D} (\mathcal{N}, \mathbf{I}),$$

where $\mathcal{N}$ is as above.

We first prove asymptotic existence and consistency. By Taylor expansion we have that

$$pl_N(\tilde{\beta} = pl_N(\beta_0) + (\tilde{\beta} - \beta_0)' S_N - \tfrac{1}{2}(\tilde{\beta} - \beta_0)' \mathbf{H}_N(\tilde{\tilde{\beta}})(\tilde{\beta} - \beta_0)$$

where $\tilde{\tilde{\beta}}$ lies between $\tilde{\beta}$ and β_0. Equivalently

$$pl_N(\tilde{\beta}) - pl_N(\beta_0) = (\tilde{\beta} - \beta_0)' S_N - \tfrac{1}{2}(\tilde{\beta} - \beta_0)' \mathbf{H}_N(\tilde{\tilde{\beta}})(\tilde{\beta} - \beta_0). \tag{22}$$

Let now $\tilde{\lambda} = \mathbf{F}_N^{t/2}(\tilde{\beta} - \beta_0)/\delta$. Then it follows that $(\tilde{\beta} - \beta_0)' = \tilde{\lambda}' \mathbf{F}_N^{-1/2} \delta$ and $\tilde{\lambda}'\tilde{\lambda} = 1$. In this notation (22) becomes

$$pl_N(\tilde{\beta}) - pl_N(\beta_0) = \delta \tilde{\lambda}' \mathbf{F}_N^{-1/2} S_N - \frac{\delta^2}{2} \tilde{\lambda}' \mathbf{F}_N^{-1/2} \mathbf{H}_N(\tilde{\tilde{\beta}})\, \mathbf{F}_N^{-t/2} \tilde{\lambda}. \tag{23}$$

We are going to prove that for every $\eta > 0$ there exist N and δ such that

$$P[\, pl_N(\tilde{\beta}) - pl_N(\beta_0) < 0 \ \forall \tilde{\beta} \in \partial O_N(\delta)\,] \geqslant 1 - \eta. \tag{24}$$

This shows that, with probability tending to one, there exists a local maximum inside $O_N(\delta)$. From (23), we recognize that it is sufficient to show

$$P\left[\ \|\mathbf{F}_N^{-1/2}S_N\|^2 \leqslant \delta^2\,\frac{\lambda_{\min}^2(\mathbf{F}_N^{-1/2}\mathbf{H}_{\mathbf{N}}(\tilde{\tilde{\beta}})\,\mathbf{F}_N^{-t/2})}{4}\right] \geqslant 1-\eta. \tag{25}$$

This is so because of the inequality

$$\begin{aligned}\tilde{\lambda}\mathbf{F}_N^{-1/2}S_N - \frac{\delta}{2}\,\tilde{\lambda}\mathbf{F}_N^{-1/2}\mathbf{H}_N(\tilde{\tilde{\beta}})\,\mathbf{F}_N^{-t/2}\tilde{\lambda}\\ \leqslant \|\mathbf{F}_N^{-1/2}S_N\| - \frac{\delta}{2}\,\lambda_{\min}(\mathbf{F}_N^{-1/2}\mathbf{H}_N(\tilde{\tilde{\beta}})\,\mathbf{F}_N^{-t/2}).\end{aligned}$$

Taking into account (15) and Lemma 3.3, we have that for sufficiently large N,

$$\begin{aligned}P\left[\ \|\mathbf{F}_N^{-1/2}S_N\|^2 \leqslant \delta^2\,\frac{\lambda_{\min}^2(\mathbf{F}_N^{-1/2}\mathbf{H}_N(\tilde{\tilde{\beta}})\,\mathbf{F}_N^{-t/2})}{4}\right] &\geqslant P\left[\ \|\mathbf{F}_N^{-1/2}S_N\|^2 \leqslant \frac{\delta^2}{4}\right] - \frac{\eta}{2}\\ &\geqslant 1 - \frac{4E[\,\|\mathbf{F}_N^{-1/2}S_N\|^2\,]}{\delta^2} - \frac{\eta}{2}\end{aligned}$$

by using Lemma 3.4. Since $E[\,\|\mathbf{F}_N^{-1/2}S_N\|^2\,] = p$, the above expression can become arbitrarily small. Asymptotic existence therefore was established. More specifically, we have that there exists a sequence $\{\hat{\beta}_N\}$ of MPLE's such that for any $\eta > 0$, there is a δ, N_1 with

$$P[\,\hat{\beta}_N \in O_N(\delta)\,] \geqslant 1-\eta \quad \forall N \geqslant N_1. \tag{26}$$

From Lemmas 3.1 and 3.3 we obtain that $\mathbf{H}_N(\beta)$ is positive definite throughout $O_N(\delta)$ with probability converging to 1. Therefore the MPLE $\hat{\beta}_N$ is also locally unique. Consistency was established as well, upon noting

$$\begin{aligned}1-\eta &\leqslant P[\,\|\mathbf{F}_N^{t/2}(\hat{\beta}_N - \beta_0)\| \leqslant \delta\,]\\ &\leqslant P\left[\ \|\hat{\beta}_N - \beta_0\| \leqslant \frac{\delta}{\lambda_{\min}(\mathbf{F}_N)}\right].\end{aligned}$$

We prove now asymptotic normality. By Taylor expansion around $\hat{\beta}_N$, and using the mean value theorem for multivariate function we obtain

$$S_N = \tilde{\mathbf{H}}_N(\hat{\beta}_N - \beta_0), \tag{27}$$

where $\tilde{\mathbf{H}}_N = \int_0^1 \mathbf{H}_N(\beta_0 + s(\hat{\beta}_N - \beta_0))\,ds$ and the integration is taken elementwise. We need to show that

$$\mathbf{F}_N^{-1/2}\tilde{\mathbf{H}}_N\mathbf{F}_N^{-t/2} \xrightarrow{\ p\ } \mathbf{I}. \tag{28}$$

But

$$\mathbf{F}_N^{-1/2}\tilde{\mathbf{H}}_N\mathbf{F}_N^{-t/2} = \mathbf{F}_N^{-1/2}(\tilde{\mathbf{H}}_N - \mathbf{G}_N)\,\mathbf{F}_2^{-t/2} + \mathbf{F}_N^{-1/2}\mathbf{G}_N\mathbf{F}_N^{-t/2}$$
$$\xrightarrow{p} \mathbf{0} + \mathbf{I} = \mathbf{I}.$$

This is so because for $N \to \infty$, the MPLE is consistent so that $\tilde{\mathbf{H}}_N - \mathbf{G}_N$ behaves the same as $\mathbf{H}_N - \mathbf{G}_N$. Invoking Lemmas 3.1 and 3.3 we have that (28) holds. Therefore, from (27)

$$\mathbf{F}_N^{-1/2}S_N = (\mathbf{F}_N^{-1/2}\tilde{\mathbf{H}}_N\mathbf{F}_N^{-t/2})(\mathbf{F}_N^{t/2}(\hat{\beta}_N - \beta_0)).$$

Thus

$$\mathbf{F}_N^{t/2}(\hat{\beta} - \beta_0) \to \mathcal{N}.$$

But

$$\mathbf{G}_N^{t/2}(\hat{\beta} - \beta_0) = \mathbf{G}_N^{t/2}\mathbf{F}_N^{-t/2}\mathbf{F}_2^{t/2}(\hat{\beta}_N - \beta_0) \to \mathcal{N}$$

since $\mathbf{G}_N^{t/2}\mathbf{N}_N^{-t/2} \xrightarrow{p} \mathbf{I}$. From the continuity of the square root.

$$\frac{\mathbf{G}_N^{t/2}}{\sqrt{N}} \xrightarrow{p} \mathbf{G}_N^{t/2}.$$

An application of Slutsky's theorem yields to the conclusion of the theorem. ∎

Remark. It will be helpful to make a comparison with Wong's important work (Wong [32]) on the general theory of partial likelihood and the results that we obtained in this work. We need to point out that in (Wong [32]), the author did not discuss explicitly models for categorical time series. However, he indicated (Wong [32, p. 96]) how the method of partial likelihood can be used to model non-stationary and non-normal time series through the broad class of generalized linear models. Wong introduced the term "generalized autoregressive" for such models. He gave sufficient conditions for the proof of consistency. The main idea was to use the accumulated Kullback–Leibler information. Our approach to the consistency problem is based on assumptions A.2 and A.5. These assumptions actually preserve Wong's spirit since we prove equation (24) which leads to the same result. However, for the proof of asymptotic normality we do not make any assumptions on the third derivatives of the partial loglikelihood. Furthermore, assumption A.5 and the remarks following it show that one needs convergence of the conditional information matrix as in Wong [32, Thm. 4A]. The Lindeberg's condition is immediately satisfied by the compactness assumption A.2.

COROLLARY 3.1. *Under* **(A)** *we have*

$$\sqrt{N}(\hat{\beta}_N - \beta_0) - \frac{1}{N}\mathbf{G}^{-1}S_N \xrightarrow{p} 0.$$

Proof. Using again Slutsky's theorem and the continuity of the square root we obtain that

$$\frac{1}{N}\mathbf{G}^{-1}S_N \xrightarrow{D} \mathcal{N}(0, \mathbf{G}^{-1}).$$

The claim follows from the previous theorem and Slutsky's theorem once again. ▮

Now, assume that each component of the link function is log-concave, that is, $\log h_j$ is concave for every $j = 1, ..., m$ with $h_m = 1 - \sum_{j=1}^{q} h_j$ and the parameter space B is R^p. Then we obtain the following:

COROLLARY 3.2. *Suppose* **(A)** *holds. Assume further that* $\log h_j$ *is concave for* $j = 1, ..., m$. *Then the probability that a unique maximum partial likelihood estimator exists converges to one. Any such sequence is consistent and asymptotically normal as in Theorem* 3.1.

4. GOODNESS OF FIT STATISTICS

A question which arises naturally after every procedure involving regression is that of goodness of fit. Our approach is to classify the responses y_t according to mutually exclusive events in terms of the covariates $\mathbf{Z}_{t-1}$ (see Schoenfeld [26]; Slud and Kedem [29]). Suppose that $A_1, ..., A_k$ constitute a partition of $R^{p \times q}$. For $l = 1, ..., k$ define

$$M_l = \sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1} \in A_l]} y_t$$

and

$$E_l(\beta) = \sum_{t=1}^{N} \mathbf{I}_{[\mathbf{Z}_{t-1} \in A_l]} p_t(\beta),$$

where I is the indicator of the set $\{\mathbf{Z}_{t-1} \in A_l\}$, for $l = 1, ..., k$. Let $M_N = (M_1', ..., M_k')'$, $E_N(\beta) = (E_1'(\beta), ..., E_k'(\beta))'$. If we let $I_{t-1} = (I_{[\mathbf{Z}_{t-1} \in A_1]}, ..., I_{[\mathbf{Z}_{t-1} \in A_k]})'$ we can see that

$$d_N(\beta) = M_N - E_N(\beta) = \sum_{t=1}^{N} I_{t-1} \otimes (y_t - p_t(\beta)),$$

where $\otimes$ denotes Kronecker product. It follows that $d_N(\beta)$ is a zero mean square integrable martingale that satisfies all the conditions needed for an application of the central limit theorem under our previous assumptions. The details of the proof are omitted. Thus

$$\frac{d_N}{\sqrt{N}} \xrightarrow{p} \mathcal{N}(0, \mathbf{C}),$$

where $\mathbf{C} = \oplus_{l=1}^{k} \mathbf{C}_l$, the direct sum of k matrices,[1] and $\mathbf{C}_l$ is a $q \times q$ symmetric matrix given by

$$\mathbf{C}_l(\beta_0) = \begin{bmatrix} \int_{A_l} p_1(\beta_0)(1 - p_1(\beta_0))\, \mu(d\mathbf{Z}) & \cdots & -\int_{A_l} p_1(\beta_0)\, p_q(\beta_0)\, \mu(d\mathbf{Z}) \\ \vdots & \ddots & \vdots \\ -\int_{A_l} p_1(\beta_0)\, p_q(\beta_0)\, \mu(d\mathbf{Z}) & \cdots & \int_{A_l} p_q(\beta_0)(1 - p_q(\beta_0))\, \mu(d\mathbf{Z}) \end{bmatrix}.$$

From the above result we have the following proposition:

PROPOSITION 4.1. *As $N \to \infty$, the asymptotic distribution of the statistic*

$$\chi^2(\beta_0) = \frac{1}{N} \sum_{l=1}^{k} d_l'(\beta_0)\, \mathbf{C}_l^{-1}(\beta_0)\, d_l(\beta_0) \tag{29}$$

is chi-square with kq degrees of freedom.

We are going to demonstrate now another theorem which gives rise to another goodness of fit statistic.

THEOREM 4.1. *Suppose that* **(A)** *hold. Let $A_1, ..., A_k$ be a partition of $R^{p \times q}$. Then we have as $N \to \infty$*

1. $\sqrt{N}\left(\frac{d_N'}{N}, (\hat{\beta}_N - \beta_0)\right) \xrightarrow{D} \mathcal{N}(0, \mathbf{\Gamma}),$

[1] $\mathbf{A} \oplus \mathbf{B}$ creates a partitioned diagonal matrix, having $\mathbf{A}$, $\mathbf{B}$ on the main diagonal.

where Γ is a square matrix of dimension $p+kq$

$$\Gamma = \begin{bmatrix} \mathbf{C} & \mathbf{B}' \\ \mathbf{B} & \mathbf{G}^{-1} \end{bmatrix}.$$

Here $\mathbf{C}$ *is as in Proposition* (29), $\mathbf{G}$ *is the limiting* $p \times p$ *information matrix, and the lth column of* $\mathbf{B}$ *is given by the matrix*

$$\mathbf{G}^{-1} \int_{A_l} \mathbf{Z}\mathbf{D}\boldsymbol{\Sigma}\mu(d\mathbf{Z}).$$

2. *We also have, as* $N \to \infty$, *that*

$$\frac{E_N(\hat{\beta}_N) - E_N(\beta_0)}{\sqrt{N}} - \sqrt{N}\,\mathbf{B}'\mathbf{G}(\hat{\beta}_N - \beta_0) \xrightarrow{p} 0.$$

Proof. For proving (1) we only need to observe from Corollary 3.1 that for some integer N greater than N_0 we have that

$$\frac{1}{\sqrt{N}}(d'_N, (\tilde{\beta} - \beta_0)) \overset{p}{\approx} \frac{1}{\sqrt{N}}(d'_N, \mathbf{G}^{-1}S_N). \tag{30}$$

Now we know that d_N and S_N are martingales which obey the conditions for an application of a central limit theorem for martingales. It follows that jointly (using again the Cramer–Wold device) the vector on the right hand side of the above equation converges to normal as $N \to \infty$. We only need to compute the asymptotic covariance matrix of its component. We have

$$\frac{1}{N}\mathbf{G}^{-1}S_N \sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1} \in A_l]}(y_t - p_t)$$

$$= \frac{1}{N}\mathbf{G}^{-1} \sum_{s=1}^{N} \mathbf{Z}_{s-1}\mathbf{D}_{s-1}(y_s - p_s) \sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1} \in A_l]}(y_t - p_t).$$

But for $s < t$

$$E[\,\mathbf{Z}_{s-1}\mathbf{D}_{s-1}(y_s - p_s)\, I_{[\mathbf{Z}_{t-1} \in A_l]}(y_t - p_t)\,]$$

$$= E[\,\mathbf{Z}_{s-1}\mathbf{D}_{s-1}(y_s - p_s)\, I_{[\mathbf{Z}_{t-1} \in A_l]} E[(y_t - p_t) \parallel \mathcal{F}_{t-1}]\,] = 0.$$

Therefore, we have from assumption A.5 that

$$E\left[\frac{1}{N}\mathbf{G}^{-1}S_N\sum_{t=1}^{N}I_{[\mathbf{Z}_{t-1}\in A_l]}(y_t-p_t)\right]$$
$$=E\left[\frac{1}{N}\mathbf{G}^{-1}\sum_{s=1}^{N}\mathbf{Z}_s\mathbf{D}_{s-1}(y_s-p_s)\sum_{t=1}^{N}I_{[\mathbf{Z}_{t-1}\in A_l]}(y_t-p_t)\right]$$
$$=E\left[\sum_{t=1}^{N}I_{[\mathbf{Z}_{t-1}\in A_l]}\mathbf{Z}_{t-1}\mathbf{D}_{t-1}\boldsymbol{\Sigma}_t\right]\xrightarrow{p}\mathbf{G}^{-1}\int_{A_l}\mathbf{Z}\mathbf{D}\boldsymbol{\Sigma}_{\mu d}(\mathbf{Z}).$$

The first part of the theorem follows. For proving the second part, we have by Taylor's expansion

$$E_N^l(\beta_N)\approx E_N^l(\beta_0)+\left[\frac{\partial E_N^l(\beta)}{\partial\beta}\right]_{\beta_0}(\hat{\beta}_N-\beta_0)+\circ_p(\|\hat{\beta}_N-\beta_0\|)$$
$$=E_N^l(\beta_0)+\left[\sum_{t=1}^{N}I_{[\mathbf{Z}_{t-1}\in A_l]}\frac{\partial p_t}{\partial\beta}\right](\hat{\beta}_N-\beta_0)+\circ_p(\|\hat{\beta}_N-\beta_0\|)$$
$$=E_N^l(\beta_0)+\left[\sum_{t=1}^{N}\mathbf{Z}_{t-1}I_{[\mathbf{Z}_{t-1}\in A_l]}\frac{\partial p_t}{\partial\gamma_{t-1}}\right](\hat{\beta}_N-\beta_0)+(\|\hat{\beta}_N-\beta_0\|)$$
$$=E_N^l(\beta_0)+\left[\sum_{t=1}^{N}\mathbf{Z}_{t-1}I_{[\mathbf{Z}_{t-1}\in A_l]}\frac{\partial p_t}{\partial l}\frac{\partial l}{\partial p_t}\frac{\partial p_t}{\partial\gamma_{t-1}}\right](\hat{\beta}_N-\beta_0)$$
$$+\circ_p(\|\hat{\beta}_N-\beta_0\|)$$
$$=E_N^l(\beta_0)+\left[\sum_{t=1}^{N}I_{[\mathbf{Z}_{t-1}\in A_l]}\mathbf{Z}_t\mathbf{D}_{t-1}\boldsymbol{\Sigma}_t\right](\hat{\beta}_N-\beta_0)+\circ_p(\|\hat{\beta}_N-\beta_0\|),$$

where l is the logits function and $\gamma_{t-1}=\mathbf{Z}'_{t-1}\beta$. So the desired result follows. ▮

Remark. From the second part of the Theorem 4.1 we obtain that

$$\frac{1}{N}(M_N-E_N(\hat{\beta}_N))=\frac{1}{N}(M_N-E_N(\beta_0)+E_N(\beta_0)-E_N(\hat{\beta}))$$
$$\approx\frac{1}{N}(M_N-E_N(\beta_0))-\sqrt{N}\,\mathbf{B}'\mathbf{G}(\hat{\beta}_N-\beta_0).$$

It follows that the asymptotic covariance matrix of $(M_N-E_N(\hat{\beta}_N))/N$ is given by $\mathbf{C}-\mathbf{B}'\mathbf{G}\mathbf{B}$. So another useful statistic is

$$\frac{1}{N}(M_N-E_N(\hat{\beta}_N))'\,[\mathbf{C}-\mathbf{B}'\mathbf{G}\mathbf{B}]^{-1}\,(M_N-E_N(\hat{\beta})),$$

where the inverse is a symmetric generalized inverse. The asymptotic distribution of this statistic is again chi-square but the number of degrees of freedom is less than or equal to $kq - 1$.

ACKNOWLEDGMENTS

The authors are thankful to an associate editor and a referee for their helpful comments.

REFERENCES

1. A. Agresti, "Categorical Data Analysis," Wiley, New York, 1990.
2. P. K. Andersen and R. D. Gill, Cox's regression models for counting process: a large sample approach, *Ann. Statist.* **10** (1982), 1100–1120.
3. E. Arjas and P. Haara, A logistic regression model for hazard: Asymptotic results, *Scand. J. Statist.* **14** (1987), 1–18.
4. P. Billingsley, "Statistical Inference for Markov Processes," Univ. of Chicago Press, Chicago, 1961.
5. E. G. Bonney, Logistic regression for dependent binary observations., *Biometrics* **43** (1987), 951–973.
6. G. E. P. Box and G. M. Jenkins, "Time Series Analysis: Forecasting and Control," 2nd ed., Holden–Day, San Francisco, 1976.
7. D. R. Brillinger, An analysis of ordinal-valued time series, *in* "Athens Conference on Applied Probability and Time Series Analysis, Vol. II: Time Series Analysis" Lecture Notes in Statist., Vol. 115, pp. 73–87 Springer-Verlag, New York, 1996.
8. D. R. Cox, Partial likelihood, *Biometrika* **62** (1975), 69–76.
9. J. P. Diggle, K.-Y. Liang, and L. S. Zeger, "Analysis of Longitudinal Data," Oxford Univ. Press, New York, 1994.
10. L. Fahrmeir and H. Kaufmann, Regression models for nonstationary categorical time series, *Time Series Anal.* **8** (1987), 147–160.
11. L. Fahrmeir and G. Tutz, "Multivariate Statistical Modelling Based on Generalized Linear Models," Springer-Verlag, New York, 1994.
12. S. J. Haberman, "The Analysis of Frequency Data," Univ. of Chicago Press, Chicago, 1974.
13. P. Hall and C. C. Heyde, "Martingale Limit Theorems and Its Applications," Academic Press, New York, 1980.
14. H. Kaufmann, Regression models for nonstationary time series: Asymptotic estimation theory, *An. Statist.* **15** (1987), 79–98.
15. H. Kaufmann, On existence and uniqueness of maximum likelihood estimates in quantal and ordinal response models, *Metrika* **13** (1989), 291–313.
16. B. Kedem, "Time Series Analysis by Higher Order Crossings," IEEE Press, New York, 1994.
17. D. M. Keenan, A time series analysis of binary data, *J. Amer. Statist. Assoc.* **77** (1982), 816–821.
18. E. L. Korn and L. S. Whittemore, Methods for analyzing panel studies of acute health effects of air pollution, *Biometrics* **35** (1979), 795–802.
19. T. Z. Lai and C. Z. Wei, Least squares estimation in stochastic regression models with applications to identification and control of dynamic systems, *Ann. Statist.* **10** (1982), 154–166.

20. W. K. Li, Time series models based on generalized linear models: some further results, *Biometrics* **50** (1994), 506–511.
21. K.-Y. Liang and S. L. Zeger, A class of logistic regression models for multivariate binary time series, *J. Amer. Statist. Assoc.* **84** (1989), 447–451.
22. P. McCullag and J. A. Nelder, "Generalized Linear Models," 2nd ed., Chapman and Hall, London, 1989.
23. L. R. Muenz and L. V. Rubinstein, Markov models for covariate dependence of binary sequences, *Biometrics* **41** (1985), 91–101.
24. S. Murphy and B. Li, Projected partial likelihood and its application to longidutinal data, *Biometrika* **82** (1995), 399–406.
25. W. J. Pratt, Concavity of the log-likelihood, *J. Amer. Statist. Assoc.* **76** (1981), 103–106.
26. D. Schoenfeld, Chi-square goodness-of-fit test for the proportional hazards regression model *Biometrika* **67** (1980), 145–153.
27. M. J. Silvapulle, On the existence of maximum likelihood estimates for the bionomial response models, *J. Royal Statist. Soc.* **B43** (1981), 310–313.
28. E. Slud, Partial likelihood for continuous time stochastic processes, *Scand. J. Statist.* **19** (1992), 97–109.
29. E. Slud and B. Kedem, Partial likelihood analysis of logistic regression and autoregression, *Statistica Sinica* **4** (1994), 89–106.
30. R. D. Stern and R. Coe, A model fitting analysis of daily rainfall data, *J. Royal Statist. Soc.* **A47** (1984), 1–34.
31. R. W.M. Wedderburn, On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models, *Biometrika* **63** (1976), 27–32.
32. W. H. Wong, Theory of partial likelihood, *Ann. Statist.* **14** (1986), 88–123.